Evaluating Seed Selection for Information Diffusion in Mobile Social Networks

Farouk Mezghani, Manel Mezghani, Ahmad Kaouk, André-Luc Beylot, Florence Sèdes

Evaluating Seed Selection for Information Diffusion in Mobile Social Networks

Farouk Mezghani[1], Manel Mezghani[2], Ahmad Kaouk[2], André-Luc Beylot[1], Florence Sèdes[2]
[1] Université de Toulouse, IRIT, INP-ENSEEIHT, France
[2] Université de Toulouse, IRIT, UPS, France

***Abstract*—The integration of social networks with mobile communication has led to the rise of a new paradigm, the mobile social network (MSN). Recently, MSN has emerged as a new hot spot of research attracting much interest from both academia and industrial sectors. For instance, MSN opens new horizon for information diffusion-based applications such as viral marketing. Thus, it is a fundamental issue to select an efficient subset of seed-nodes (i.e. initial sources) in a MSN such that targeting them initially will maximize the information diffusion to interested nodes. This paper studies the problem of identifying the best seeds through whom the information can be diffused in the network in order to maximize the content utility (i.e. a quantitative metric that determines how satisfied are the users). A multi-layer model that combines the social relationships and the mobile network in order to design an efficient information diffusion is proposed. Based on this multi-layer model, different seed selection approaches are proposed for information diffusion environment (e.g. mobile advertising) where users have heterogeneous interests for the different information generated in the network. Simulation results show the effectiveness of multi-layer based seed selection approaches comparing to a classical approach.**

I. Introduction

The rapid proliferation of mobile devices (e.g. smartphones, tablets) had created a new channel for mobile applications. Mobile devices are nowadays equipped with several features such as: multiple network interfaces for long-range (e.g. cellular) and short-range (e.g. WiFi) communications, Global Positioning System (GPS), and others.

Online social networks (OSN) such as Facebook and Twitter have exploited these aforementioned features and have extended their services to mobile devices through mobile applications to mobilize their networks and thus to provide more advanced social services. Hence, the integration of social networks with mobile communication networks has led to the rise of a new paradigm, the mobile social networks (MSNs) [1].

MSNs have opened new opportunities for information diffusion based application such as viral marketing and mobile advertising [2–4] which relies on spreading information for mobile users with exploiting their social relationships. Diffusing information for all target-users (i.e. interested users) using only cellular communication is not efficient for different reasons. Indeed, it has an expensive cost for service providers, it might overload cellular networks, and users may not have or have a limited access to cellular network due to its high cost. Therefore, diffusing information only through a set of users called seed-nodes presents a promising solution [5]. Indeed, seed-nodes can download information from content server through cellular network, then, they diffuse the information to proximity users using low-cost short-range communication such as WiFi. However, many challenges need to be addressed before fully exploiting MSN as a platform for information diffusion. For instance, the design of an efficient information diffusion should address two main challenges: *which node to select as seed?* and *how to minimize the number of seeds?*.

Most of existing information diffusion approaches [3], [6–8] target to optimize the number of seeds and/or to select the seeds allowing to rapidly reach a maximum number of users. Basically these approaches exploit criteria such as social relations between nodes, contacts history and node mobility to select the seed-nodes. However, some of the picture is still missing an important feature: *user interests*. Indeed, a large amount of content might be generated in the network while a user is seldom interested in all these content. Naturally, a user wants only a small part of these information according to its personal preferences [9]. Thus, it is not efficient to diffuse information to all users (i.e. blind diffusion) in the network since it might not satisfy their interests.

The goal of this paper is thus to design an information diffusion that considers heterogeneous users' preferences and targets to maximize the content utility (i.e. a quantitative metric to determine how much users are satisfied). Therefore, an interest-based multi-layer model is proposed to facilitate the design of information diffusion. Then, three seed selection approaches, based on the multi-layer model, are proposed and evaluated through extensive simulations. These approaches are compared with a classical seed selection considered as a reference approach.

The paper proceeds as follows: Section II gives an overview of the related works. Section III presents the system model and details the problem statement. Section IV proposes different seed selection approaches based on a novel multi-layer model for information diffusing in MSN. In Section V, proposed approaches are evaluated through intensive simulations and compared with a reference approach. Section VI concludes the paper and gives some perspectives.

II. Related Works

A. Mobile Networks come to social

In social networks, data are usually represented as a social graph. It describes the structure of the social network. In a social graph, nodes represent humans or mobile devices, while the edges describe the relationships between these individuals [10]. Several types of social graphs can be defined according to the representation of the links between the nodes. For example, a social graph can be viewed as a contact graph where the links describe the contact between nodes. Such a graph is very popular and used in analyzing and estimating the relationships among nodes. There are also other types of social graphs such as interest graph and the neighbor graph.

In this social graph, nodes can be grouped together and form communities. A community is a sub unit of the social graph with a high density of internal social links [11]. Nodes in the same community being more willing to share information and resources than with nodes outside. They may contact or share information with higher probability. Community-based approaches aim to adapt information in social networks in order to satisfy more users [12].

Many research activities have studied social properties of mobile nodes in mobile social networks to design more efficient network protocols [11]. There are many social properties that can be utilized to measure the relationships of nodes in MSN such as centrality, similarity, tie strength, etc. This work focus on the centrality metric which is one of the most popular social metrics used in MSNs to analyze social behavior and relationships. The centrality metric describes the importance of a nodes in a social graph. It is widely used to describe nodes and groups to identify most important and prominent actors. A central node has a stronger capacity of connecting other nodes [13]. Several ways exist to measure the centrality of nodes, from which the most popular and widely used are: betweeness, closeness and degree centrality.

B. Information Diffusion in MSN

MSN is a fertile ground for information spreading. The process of information diffusion in MSN has been widely used especially for viral marketing [7]. In general, the process of information diffusion is comprised of two major parts [4]: *seed selection* which represents the focus of this study and *information diffusion between nodes* which defines the opportunistic exchange between nodes.

In recent studies, seed selection is operated at the goal of maximizing information spreading ratio and/or minimizing the diffusion time. Given limited resources for hiring seed-nodes, selecting a minimal number of nodes to achieve maximum cover with a given period of time is NP-Hard [14]. Therefore, heuristic solutions have been proposed.

A contact strength-based seed selection scheme is proposed and used for viral marketing [7] and aims to maximize the information propagation in the network. This work selects as seeds the nodes that have good social strength with other nodes in the network computed based on contact history. In [3], [6], authors propose a community-based algorithm in order to minimize the diffusion time. It considers node centrality and contact frequency to select as seeds the most influential nodes inside a community. This work addresses the problem of identifying a small number of individuals through whom the information can be diffused to the network as soon as possible. Another community-based seed selection solution [8] have been proposed to select seeds as the nodes having the highest centrality but belonging to different communities. Thus, seeds are selected in different communities and can then reach a maximum number of users.

The aforementioned cited works achieved important results in terms of information propagation and delivery delay. However, an important feature has been left behind : *users interests*. Indeed, these works assume that all participating nodes are interested in the generated information. Basically, users may be interested in receiving all critical information (e.g. terrorist attack alert) since their life is at sake. But, for other types of information (e.g. restaurant recommendation, theater offers), users are seldom interested in the large amount of content that can be generated everyday. They only want a specific useful information based on their personal preferences.

Therefore, unlike existing works, in this study, heterogeneous user preferences are considered to design seed selection that aims to maximize users satisfaction with respect to received content.

III. System Model and Problem Statement

A. System Model

This work considers a MSN system model as illustrated in Figure 1. The network system is composed of mobile users $\mathcal{U} = \{u_i\}$ having mobile devices (e.g. smartphones) equipped each with two different network interfaces: a cellular interface and a short range communication interface.

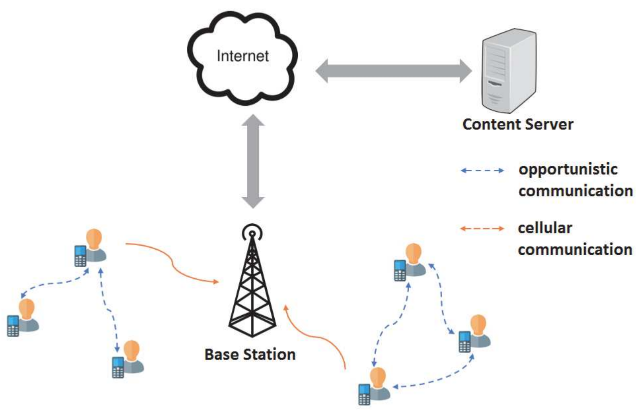


Fig. 1. Network system

The system also involves an Internet-content server from which seed-nodes (i.e. initial sources) retrieve content objects initially through the cellular networks. Seeds can then exploit the direct communications to diffuse downloaded content objects to neighboring nodes.

This work focuses on information diffusion applications such those used by service providers that want to announce users about their services and attract them to their stores (e.g. announcements about shopping center discount, restaurants recommendation, theater offers and others). The term *content* *object* represents a specific information, for instance, regarding a specific advertisement, such as a restaurant advertisement.

Each user u_i is required to explicitly provide its preferences for different predefined topics $\mathcal{T} = \{T_k\}$ (e.g. theater offers, restaurant recommendation). Preferences are represented in order to express different interest level, e.g. not interested, a little interested, interested, and very interested. Indeed, nowadays, users have heterogeneous interests for different topics (e.g. a user can have more interest for sport announcement than for a theater offer). Let I_{i,T_k} and $I_{i,j}$ denote the interests of the user u_i for the topic T_k and the content object o_j, respectively. $I_{i,j} = I_{i,T_k}$ if o_j is a content object regarding the topic T_k. Hence, when a user u_i receives a content object o_j regarding a topic T_k, s/he gets the $utility = I_{i,j} = I_{i,T_k}$ where $o_j \in T_k$.

This study considers "*Content Utility rate*", a quantitative metric that computes the nodes gain by means of interests (i.e. how satisfied the users are) as shown in equation 1, to evaluate the information diffusion [9]. An efficient information diffusion should then produce a maximum *content utility rate*.

$$Content\ Utility\ rate = \frac{\sum_{u_i \text{ received } o_j} I_{i,j}}{\sum I_{i,j}} \tag{1}$$

B. Problem Statement

Let's assume o_z a new content object created at instant t and available within the content server. For o_z diffusion and in order to increase the users satisfaction, the efficient seeds $\mathcal{S} = \{s_n\}$ for o_z are the nodes that can reach a maximum number of interested users and can achieve a maximum content utility. This problem presents an NP-hard problem [14]. In the following, heuristic solutions are proposed and evaluated in order to select preferable seeds targeting to maximally satisfy users interests.

IV. Seed Selection in MSN - What information to exploit ?

This section presents a novel interest-based multi-layer model for information diffusion in MSN. Then, multi-layer based seed selection approaches, that target to maximally satisfy users interests, are proposed.

A. Interest-based multi-layer model

Figure 2 overviews the system model composed of *multiple topic layers* and *network layer*. Topic layers are defined based on the list of the predefined topics $\mathcal{T}$ and users interests. Each layer T_k-*layer* provides information only about users interested in its corresponding topic T_k (i.e. users that have interest in T_k: $I_{i,T_k} > 0$). Let $\mathcal{U}_k = \{u_i \in T_k\text{-}layer\}$ denote the users that belong to the T_k-*layer*. A user can belong to multiple layers. The network layer represents the services of the physical, logical, and the routing layers of the protocol stack.

We note in real environment, a global view of the network can be achieved using a centralized equipment. Indeed, this equipment collects periodically information about users (i.e. interests, location) through a real-time location tracking system. Thus, these information can be exploited to build the multi-layer model offering a simple view of the networks.

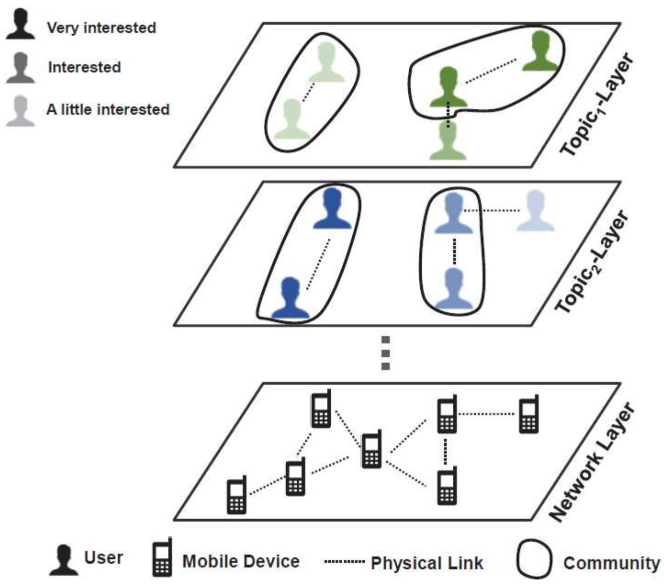


Fig. 2. Multi-layer model

1) Network layer: it considers all mobile users in the network. Each user has a mobile device allowing it to communicate directly with other users in proximity.

2) Topic layer: On each layer, heterogeneous preferences are considered and presented with graduated color (i.e. more the color is dark more the user is interested in the layer topic) as illustrated in Figure 2.

When a new content object regarding a topic T_k is created, it is then diffused considering information on the T_k-*layer*. Seed-nodes are then selected from nodes on the T_k-*layer* to download the content object and diffuse it to their neighbors. However, these information can be exploited in different manners to select the seed-nodes. Hence, in the following, we investigate different multi-layer based seed selection solutions.

B. Multi-layer based seed selection

This section discusses the seed selection in MSN which represents the first step of information diffusion. The main goal of this part is to design an efficient seed selection allowing to maximize the users satisfaction from the content diffusion. Thus, understanding the impact of some criteria such as users interests, node centrality, community on the content diffusion is of great importance.

Three different seed selection approaches based on the multi-layer model are proposed: (1)multi layer-based centrality, (2)multi layer-based utility, and (3)multi layer based community.

1) Multi-layer based centrality: this approach considers two criteria: users binary interests (i.e. does not consider heterogeneous interests) and node centrality. In order to reach a maximum number of "interested" users, this method selects as seeds nodes having the highest centrality in the topic-layer.

This method considers degree centrality to select seeds. Degree centrality is defined as the number of one-hop neighbors of a node. Thus, for a T_k-$layer$, the degree centrality DegC of a node u_i is computed as shown in equation 2.

$$DegC(u_i) = \frac{\#\mathcal{N}(u_i)}{\#\mathcal{U}_k} \tag{2}$$

where $\mathcal{N}(u_i)$ are the directly connected neighbors of u_i and belongs to T_k-$layer$.

2) Multi-layer based utility: this method is based on users heterogeneous interests. The utility that a node can produce to its neighbors is computed knowing its neighbors and their interests as showed in equation 3.

$$Utility(u_n, o_j) = \sum_{u_i \in \mathcal{N}(u_n)} I_{i,j} \tag{3}$$

where $\mathcal{N}(u_n)$ is the direct neighbors of u_n ; $u_i, u_n \in T_k$-$layer$ and $o_j \in T_k$. Nodes that can produce the highest utilities are selected as seeds.

3) Multi-layer community-based community: it relies on the creation of communities based on social relations between users. Then it selects as seed the nodes that have a high centrality and belong to different communities. There are many state of art community detection algorithms which does not represent the focus of this work. A simple heterogeneous interest-based method is considered to form communities in the topic-layer. Indeed, in each layer, communities are formed considering users that share the same interest level as shown in Figure 2. Seeds are selected as the nodes belonging to different communities and having the highest centrality inside the community.

V. Performance Evaluation

A. Methodology

Simulations have been conducted through the Opportunistic Network Environment Simulator (The ONE) [15]. Simulations involve 250 mobile users moving with a velocity in the range of 0.5∼1.5 m/s. Each user is supposed to have a mobile device equipped with two interfaces; cellular interface and short range communication interface. Short range communication interface is considered with a transmission range of 60 m and a transmission speed of 2 Mbps. Simulation scenarios consider two different mobility models; (i) *cluster movement* where users move randomly in restricted small area and (ii) *Map-based movement* in which users follow a map-driven mobility model [16] on Helsinki downtown area [15].

Users express randomly with a uniform distribution their preferences for different predefined topics considering the interest values: 0, 1, 2, and 3 which corresponds to the different interest level: not interested, a little interested, interested, and very interested, respectively. A number of 200 content objects with equal size 1 MB is generated with an inter-object interval time in a range of 60 seconds. Each content object has a limited lifetime of 1 hour.

Simulations evaluate and compare the performance of our multi-layer seed selection proposals: (i) *Interest layer centrality*, (ii) *Community centrality*, (iii) *utility*, and *Network centrality-based* seed selection considered as a reference in literature works. Indeed, most of existing seed selection approaches are based on criteria to determine nodes playing a vital role in the network. These criteria are mainly based on the nodes centrality. These approaches have shown through simulations results that centrality-based method is a major factor giving the best performance compromise [11].

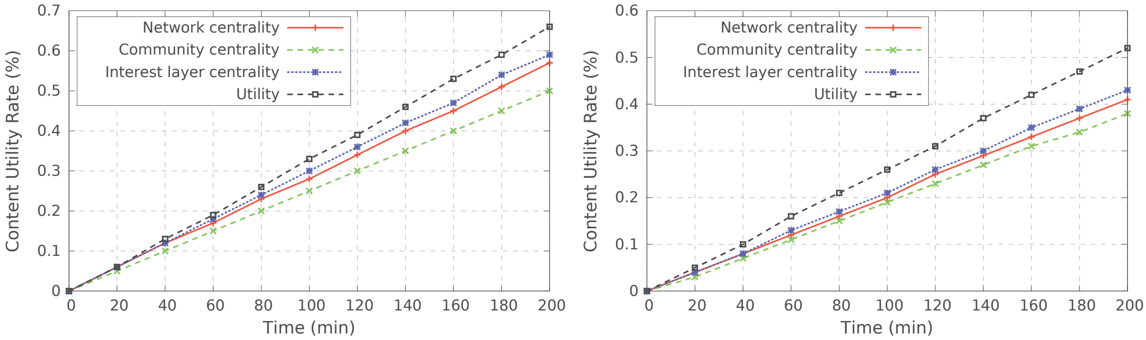
In first scenarios, 20% from all nodes in the network are selected as seeds. These latter download content objects from the content server through cellular networks. Then, seeds diffuse content objects to neighboring nodes using short range communication. In order to not overload the network, and to maximally reduce the duplicated message, it is assumed in the first scenarios that a content object is diffused only one time from each seed-node.

B. Simulations results

Figure 3 illustrates the content utility produced over time for cluster-based and map-based mobility models. In a cluster-based scenario (as shown in Figure 3a), results show that utility-based seed selection outperforms the other schemes. Indeed, in the end of the simulation, it has produced 52% of content utility rate while Interest-layer centrality, network centrality, and community centrality realized only 43%, 41% and 38%, respectively. Similarly, the map-driven scenario shows that utility-based approach performs better than the other seed selection schemes. As shown in Figure 3b, it satisfied 66% from the users interests while interest-layer centrality, network centrality, and community centrality satisfied only 59%, 57% and 50%, respectively.

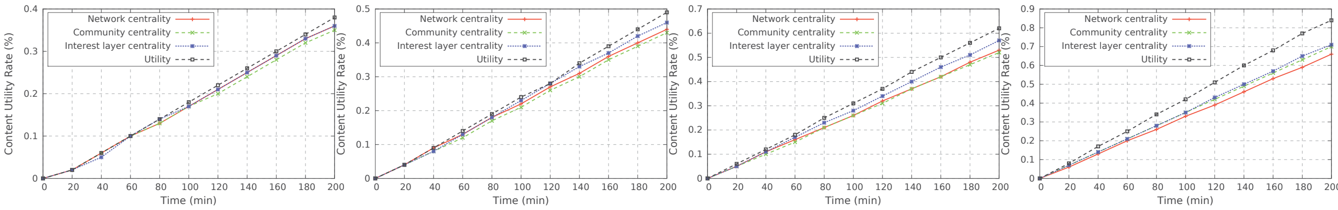
The difference of the results obtained between movement-based and map-based movement scenarios are due to the network density and nodes mobility which consequently impacts the number of contacts between nodes. In cluster movement, the density of the network is higher than in the map based movement. Consequently, nodes in cluster movement can have more contact frequency and more direct neighbors comparing to the map based network. Therefore, seeds in cluster movement can reach more nodes than in map based movement, and then achieve better results in terms of content utility rate.

Interest layer centrality and network centrality presents very similar results in both previous scenarios. Both approaches target to disseminate content to a maximum number of users by selecting as seeds the nodes that play a vital role in the network. Interest-layer centrality slightly outperforms network centrality. Indeed, network centrality select as seeds the nodes that have a higher centrality among all nodes in the network. This can lead to a diffusion for a maximum number of users which might be not interested in the received content. Instead, interest-layer centrality design as seeds the nodes that presents



(a) Cluster movement (b) Map based movement

Fig. 3. Cumulative content utility over time



(a) 4 seeds, 10 transmission per seed (b) 10 seeds, 4 transmissions per seed (c) 20 seeds, 2 transmissions per seed (d) 40 seeds, 1 transmission per seed

Fig. 4. Content utility rate in terms of seeds number and number of transmission per seed

the higher centrality among all interested nodes in the network (i.e. topic-layer users).

Simulations results demonstrate that, considering only network parameters (e.g. centrality), seed selection approach would not be efficient since it can not maximally satisfy users interests. Therefore, taking into account users parameters such as heterogeneous interests can maximize the content utility produced. Consequently, it is not important to disseminate content to as maximum number of users as possible. Instead, it is more efficient to disseminate content to the right users (interested users or most interested users) in order to maximize the content utility.

Is it better to increase the number of seeds or the transmissions per seed?

First simulation results show that considering a limited number of seeds and making only one transmission per seed, the different approaches can not satisfy all the interested users. Therefore a compromise is needed between the number of seeds and the number of transmissions. On the one hand, increasing the seeds number involves a high cost. On the other hand, a high number of transmission may lead to a duplicated content diffusion and may increase the network overhead.

Therefore, various simulations scenarios were conducted considering various seeds number and various number of transmissions per seed in order to see the impact of these parameters on the content utility produced in the network. Thus, the following scenarios consider a high number of seeds coupled with a low number of transmissions and inversely, a low number of seeds are coupled with a high number of transmissions. The following simulations consider these combinations for (#seeds, #transmissions): (4,10), (10,4), (20,2), and (40,1). The number of users are set to 100 in the following scenarios considering a map-based movement.

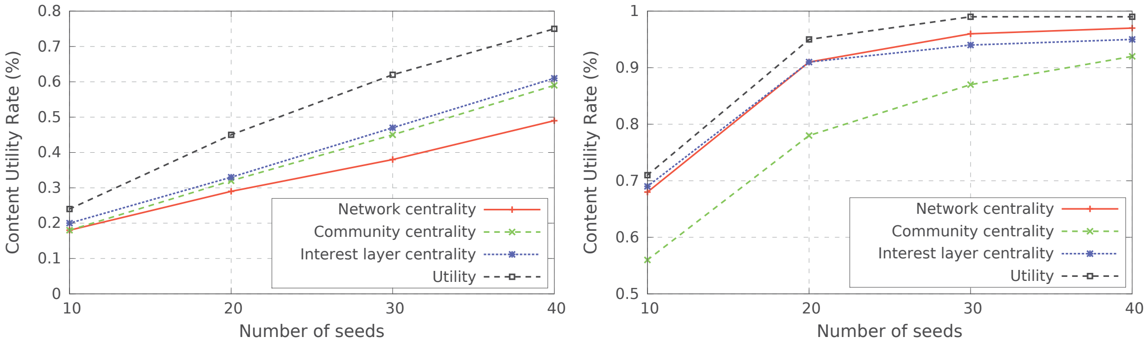
Figure 4 illustrates the content utility rate produced for the different combinations of number of seeds and number of transmissions. Results show that seed selection approaches produce more content utility when considering a high number of seeds and a low number of transmissions. Indeed, considering 40 and 20 seeds per each content object, content utility rate is higher than 60% and more than 50% as shown in Figures 4d and 4c, respectively. Differently, for a low number of seeds coupled with a high number of transmissions, Figures 4a and 4b shows that content utility is less than 40% and 50%, respectively.

In MSN, mobile users might not contact each other frequently during a short period of time. Consequently, increasing the number of transmissions will likely lead to diffuse a duplicated content and not reach other interested users. Therefore, increasing the number of seeds while reducing the number of transmissions help, on the one hand, to reach a maximum number of interested users which consequently maximize the content utility produced, and on the other hand, reduce the duplicated diffusion and do not overload the network.

How many seeds for each content object?

In the following, the impact of the number of seeds on the content utility is studied considering two different distributions of nodes positions all over the network. Only one transmission is considered per each seed-node.

First scenario considers that nodes are spread uniformly all over the network such as in a real environment, people walking in a different pedestrian zones in the Helsinki map. Second



(a) Nodes positions concentrated in different zones

(b) Nodes positions spread all over the network

Fig. 5. Cumulative content utility considering different nodes positions distribution

scenario supposes that nodes are concentrated in some specific areas such as people walking inside a shopping center or mall. Figure 5 shows that the cumulative content utility for these two scenarios considering various number of seeds 10∼40.

On the one hand, Figure 5b shows that the content utility increases when the number of seeds increases since nodes are spread uniformly all over the network. Moreover, considering only nodes centrality is not efficient to select the seeds since nodes are spread uniformly in the network and thus they present a low centrality. On the other hand, Figure 5a demonstrates that starting from a certain threshold, increasing the number of seeds does not impact the content utility. Indeed, when nodes are concentrated in some areas, this will certainly minimize the number of seeds required to reach all the interested users.

Therefore, it would be efficient to divide a big area or a region into different small zones according to the concentration of the people. Hence, an efficient content diffusion will first chose an optimal number of seeds according to the zone and the users concentration in it. Then, an efficient seed selection is realized such as to consider *layer-based utility* which can maximize the content utility.

VI. Conclusion

This work proposes a new scheme to personalize information diffusion in mobile social networks. A multi-layer model based on users interests is proposed offering a simple view and helping to design of an efficient information diffusion. Then, seed selection approaches based on the multi-layer model are proposed and target to maximally satisfy the users interests. The efficiency of the proposed schemes is emphasized through simulation studies and results showed that multi-layer based seed selection schemes allow to maximize the content utility. Determining the efficient number of seeds and number of transmissions for each content object is a focus of future work. Basically, this number should be dynamic since it is related to different features such as the content object popularity, the concentration of nodes in a zone, and the object lifetime. For instance, considering a shot object lifetime, a low number of transmission is more preferred since users usually do not meet many other nodes during a short period of time. Similarly, for a zone with a high concentration of users, the number of seeds can be minimized.

References

[1] X. Hu, T. H. Chu, V. C. Leung, E. C.-H. Ngai, P. Kruchten, and H. C. Chan, “A survey on mobile social networks: Applications, platforms, system architectures, and future research directions,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1557–1581, 2015.

[2] T. Ning, Z. Yang, H. Wu, and Z. Han, “Self-interest-driven incentives for ad dissemination in autonomous mobile social networks,” *IEEE INFOCOM*, pp. 2310–2318, 2013.

[3] Z. Lu, Y. Wen, W. Zhang, Q. Zheng, and G. Cao, “Towards information diffusion in mobile social networks,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 5, pp. 1292–1304, 2016.

[4] Q. Yao, X. Wu, and X. Zhang, “Diffusion of information in mobile social networks: A brief survey,” in *2015 IEEE International Conference on Mobile Services*. IEEE, 2015, pp. 254–260.

[5] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, “Mobile data offloading through opportunistic communications and social participation,” *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 821–834, 2012.

[6] Z. Lu, Y. Wen, and G. Cao, “Information diffusion in mobile social networks: The speed perspective,” *IEEE INFOCOM 2014*, pp. 1932–1940, 2014.

[7] B. Fan, S. Leng, K. Yang, and J. He, “Gathering point-aided viral marketing in decentralized mobile social networks,” *IEEE Systems Journal*, pp. 1–12, 2016.

[8] Y.-J. Chuang and K. C.-J. Lin, “Cellular traffic offloading through community-based opportunistic dissemination,” *IEEE WCNC*, pp. 3188–3193, 2012.

[9] F. Mezghani, R. Dhaou, M. Nogueira, and A.-L. Beylot, “Content dissemination in vehicular social networks: taxonomy and user satisfaction,” *IEEE Communications Magazine*, vol. 52, no. 12, pp. 34–40, 2014.

[10] K. Yang, X. Cheng, L. Hu, and J. Zhang, “Mobile social networks: state-of-the-art and a new vision,” *International Journal of Communication Systems*, vol. 25, no. 10, pp. 1245–1259, 2012.

[11] F. Xia, L. Liu, J. Li, J. Ma, and A. V. Vasilakos, “Socially aware networking: A survey,” *IEEE Systems Journal*, vol. 9, no. 3, pp. 904–921, 2015.

[12] M. Mezghani, C. A. Zayani, I. Amous, A. Péninou, and F. Sèdes, “Dynamic enrichment of social users’ interests,” in *IEEE RCIS*, 2014, pp. 1–11.

[13] E. M. Daly and M. Haahr, “Social network analysis for routing in disconnected delay-tolerant manets,” in *ACM MobiHoc*, 2007, pp. 32–40.

[14] J. Qin, H. Zhu, Y. Zhu, L. Lu, G. Xue, and M. Li, “Post: Exploiting dynamic sociality for mobile advertising in vehicular networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 6, pp. 1770–1782, 2016.

[15] A. Keränen, J. Ott, and T. Kärkkäinen, “The ONE Simulator for DTN Protocol Evaluation,” *SIMUTools*, p. 55, 2009.

[16] A. Keränen, T. Kärkkäinen, and J. Ott, “Simulating Mobility and DTNs with the ONE,” *Journal of Comm.*, vol. 5, no. 2, pp. 92–105, 2010.